The user manual

Index

Sir2004: General Information and Background pg. 3

Sir2004: Main Features pg. 4

Description of Sir2004 pg. 4

- Data Module pg. 4
- Invariants Module pg. 5
- Phase Module pg. 5
  - Phasing Tools pg. 6
  - The Tangent procedure pg. 7
  - The early Figure of Merit (eFOM) pg. 7
  - The RELAX procedure pg. 8
  - The Patterson deconvolution procedure pg. 9
  - The Direct Space Refinement (DSR) pg. 10
  - The molecular envelope pg. 11
  - Identification of the correct solution: the final Fom (fFOM) pg. 11

Sir2004: Strategy for *ab initio* phasing of crystal structure pg. 12

Sir2004: Completion and refinement of the structure pg. 14

- The least squares refinement pg. 14

When default Sir2004 fails: some advices pg. 15

Commands and their use pg. 21

Directives and their use pg. 23

- Data Routine pg. 23
- Invariants Routine pg. 24
- Phase Routine pg. 25

Examples of input for Sir2004 pg. 27

References pg. 32

Sir2004: General Information and Background

The **SIR** (SEMI-INVARIANTS REPRESENTATION) package has been developed for solving crystal structures by Direct Methods. The REPRESENTATION THEORY, proposed by Giacovazzo (1977, 1980) allowed the derivation of powerful methods for estimating structure invariants (s.i.) and structure seminvariants (s.s.). The mathematical approach makes full use of the space group symmetry. SIR uses symmetry in a quite general way allowing the estimation and use of s.i. and s.s. in all the space groups.

The present version of the program, **Sir2004** (Burla, Caliandro, Camalli, Carrozzini, Cascarano, De Caro, Giacovazzo, Polidori & Spagna, 2004), is designed to solve *ab initio* structures of different size and complexity, up to proteins, provided that data resolution is no lower than 1.4-1.5Å. Data can be collected with X-Ray or electron sources. There is no limit to the number of reflections and to the number of atoms in the asymmetric unit. The maximum value allowed for |h|, |k|, |l| is 512. The maximum number of different atomic species is 8.

Sir2004 includes several new features with respect to the previous version, Sir2002 (Burla, Camalli, Carrozzini, Cascarano, Giacovazzo, Polidori & Spagna, 2003). New tools are represented by the use of procedures based on Patterson Methods as alternative to the application of the Tangent Formula in order to compute the starting phase set; the introduction of suitable figures of merit (FOM's) in order to recognize the correct trial solution and the application of new algorithms for solving *ab initio* protein structures (also for quasi-atomic resolution data), i.e. the use of the molecular envelope mask.

The range of options available to experienced crystallographers for choosing their own way of solving crystal structures is rather wide. However, scientists untrained in Direct Methods or people trustful in the SIR default mode often can solve crystal structures without personal intervention.

The program is available for Microsoft Windows and Unix (SGI, Compaq, Linux and others) platforms (see the Notes on the implementation).

Authors:

M.C. Burla[(+)], R. Caliandro[(*)], M. Camalli[(*)], B. Carrozzini[(*)], G.L. Cascarano[(*)], L. De Caro[(*)], C. Giacovazzo[(*)], G. Polidori[(+)], R. Spagna[(*)].

[(*)] Istituto di Cristallografia, CNR [(+)]Dipartimento di Scienze della Terra, Univ. di Perugia.

Support:

sirmail@ic.cnr.it

Web site:

http://www.ic.cnr.it/

Sir2004: Main Features

The program has been designed to:

· require a minimal information in input;

· work automatically;

· reduce the user intervention and facilitate the interaction by means of a friendly graphic interface.

Description of Sir2004

The main modules of the program are: DATA, INVARIANTS, PHASE. The flow diagram of the program is shown in Fig. 1.

If a graphic interface is active, it is possible to interact with the model in order to complete and refine it. This option is available only if the number of atoms in the asymmetric unit does not exceed 500.

The graphic interface incorporates an online help in order to describe all the features and tools available via graphics.

Thanks are due to T. Pilati (Polyhedra representation) and to J. Gonzalez-Platas (Contouring tools) for their code integrated in Sir2004.

Data Module

This routine reads the basic crystallographic information like cell parameters, space group symbol, unit cell content and reflections. It includes a modified version of the subroutine SYMM (Burzlaff & Hountas, 1982). Symmetry operators and information necessary to identify structure invariants (estimated in INVARIANTS module) are directly derived from the space group symbol.

Diffraction data are checked in order to merge equivalent reflections, to find out systematically absent reflections (which are then excluded from the data set) and, eventually, (weak) reflections not included in the data set (Cascarano *et al*., 1991).

Diffraction intensities are normalized using the Wilson Method. Statistical analysis of intensities is made in order to check the space group correctness, to suggest the presence or absence of the inversion centre and to identify the possible presence and type of pseudotranslational symmetry (Cascarano *et al.*, 1988 a,b; Fan *et al*., 1988). Possible deviations (of displacive type) from ideal pseudotranslational symmetry are also detected. N_{large} reflections (those with the largest |E| values) are selected for invariants calculations; their maximum number is 4000.

Invariants Module

Up to 300000 triplets relating the N_{large} reflections are stored for active use in the phasing process.

Negative quartets are generated by combining the psizero triplets (relating two reflections with large |E| and one with |E| close to zero) in pairs; those with cross-magnitudes smaller than a given threshold are estimated by means of their first representation, as described by Giacovazzo (1976). These quartets are actively used in the phasing process (Giacovazzo *et al*., 1992).

Active triplets may be estimated according to Cochran's distribution (1955): the concentration parameter of the von Mises distribution is then

$$C = 2|E_{\mathbf{h}}||E_{\mathbf{k}}||E_{\mathbf{h-k}}| /\sqrt{N}$$

Triplets can also be estimated according to their second representation (i.e. P_{10} formula, as described by Cascarano *et al*., 1984). The concentration parameter of the new von Mises (i.e. of the same form of Cochran's) distribution is given by

$$G = C (1 + q)$$

where q is a function (positive or negative) of all the magnitudes in the second representation of the triplet. The G values are rescaled on the C values and the triplets are ranked in decreasing order of G. The top ranked relationships represent a better selection of triplets (with phase value close to zero) than that obtained sorting triplets according to C. Triplets estimated with a negative G value represent a sufficiently good selection of relationships close to 180 degrees. Positive and negative triplets will be actively used in the phase determination process (Giacovazzo *et al*., 1992).

The P_{10} formula is applied, as a default, to estimate triplets relationships. They are ranked in decreasing order of the concentration parameter and actively used only if G is greater than a given threshold (Default value is 0.3). If the number of triplets per reflection is too low, the program decreases the threshold to a suitable value.

Negative quartets are also generated by combining the psizero triplets (relating two reflections with large |E| and one with |E| close to zero) in pairs; those with cross-magnitudes smaller than a given threshold are estimated by means of their first representation, as described by Giacovazzo (1976). These quartets are actively used in the phasing process (Giacovazzo *et al*., 1992).

Phase Module

Before starting the phasing process, the observed |E| values of the low resolution reflections (up to 5Å) are modified (Burla *et al*., 2000). This action is undertaken since large regions of the unit cell of macromolecular crystals are filled by solvent. Since the molecular envelope is usually unknown at this stage, low resolution diffraction intensities contain an unpredictable but important

solvent contribution. Restating more reasonable values for protein diffraction intensities, by proper subtraction of the solvent effects, may be decisive for the success of the phasing process. When the structure of the macromolecule is known at high resolution, information about the solvent structure may be obtained by assuming

$$F_o = F_p + F_s$$

where F_o is the observed value for the structure factor (complex quantity), F_p is the protein structure factor calculated from the known coordinates, and F_s is the solvent contribution.

When the structure is unkown, the problem is the reverse: the aim is to obtain, at low resolution, more reasonable $|F_p|$ values from the $|F_o|$ ones, without any prior information on the crystal structure or on the molecular envelope. A simple solvent model arising from the Babinet's principle is used, i.e.,

$$F_{om} = F_p\,[1 - K \exp(-B_s r^{*2})]$$

where $r^* = 2(\sin\Theta/\lambda)$, K is a suitable scale factor (in the range 0.7-0.9) and the B_s parameter is a correction term arising from the larger vibrational motion of the solvent atoms (in the range 200-500), in accordance with Tronrud (1997) .

The expected effects of modifying the observed intensities may be so described:

a) low resolution reflections have $|F_o|$ modules too small to influence the phasing process. In the practice, very low resolution reflections are out of such process: on the contrary, the corresponding $|F_p|$ values, much larger then $|F_o|$'s (e.g., 5 times larger) could drive the phasing process along different directions.

b) Low resolution reflections are insensitive to fine structural details, but are particularly able to define the location and the envelope of the molecule. Accordingly, they play a critical role in defining the region of the electron density map selected for performing the transformation $\rho \rightarrow \varphi$, in the following steps.

Phasing Tools

According to circumstances, the Sir2004 phasing process may apply tangent procedures and/or Patterson methods; phase extension and refinement are achieved by Direct Space Refinement (DSR) techniques. To preserve its efficiency the program distinguishes different categories of structures: Xsmall molecules (up to 5 atoms per asymmetric unit), small (up to 80), medium (up to 200), large (up to 600), Xlarge (up to 1000) and XXlarge (no upper limit). While the same tangent procedure is applied to all the categories, the DSR may be accomplished in different ways according to the category (it requires increasing computing times moving from Xsmall to XXlarge structures). We shortly describe the various tools the user may employ in the phasing process.

The Tangent procedure

In Sir2004, a multisolution approach is used; the phasing procedure involves, for each trial, the application of a single tangent (ST) procedure (Burla, Carrozzini, Cascarano *et al*., 2003) to N_{large} reflections (those selected by the normalization process), starting from a subset of random phases (Baggio *et al*., 1978; Burla *et al*., 1992). Besides triplets, also the most reliable negative quartets may be actively used during the phasing process. Each relationship is used with its proper weight: the concentration parameter of the first representation G for triplets or C for quartets.

The phasing process of Sir2004 computes an early figure of merit (*eFOM*) for each tangent trial: only the best trial solutions, sorted by *eFOM,* are submitted to the DSR module. This phasing strategy allows us to explore numerous trials without paying so much in terms of computing time: it thus increases the probability of submitting to the DSR procedure a good set of phases.

In this way, we obtain the following advantages: (i) to lead to a quick solution those structures for which, owing to the good triplet estimates, the ST module frequently provides a favourable structural model; (ii) to solve those structures for which the triplet invariant estimates are so bad that a large number of trials are necessary before a good set of phases can be produced by the ST module, and subsequently submitted to the DSR procedure; (iii) to solve even those protein structures for which the *eFOM* ranking is relatively inefficient.

The early Figure of Merit (eFOM)

eFOM is defined as follows:

a) for small and medium size structures $eFOM \equiv RAT$, where (Burla *et al*., 2002)

$$\mathrm{RAT} = \frac{\mathrm{CC}_{w,R}}{\left\langle R_c^2 \right\rangle} \ .$$

$\mathrm{CC}_{w,R}$ is the correlation coefficient (in the range [0.3,1.2]) between the R_o's (say $R_o = |E_o|$)and the corresponding Sim-like coefficients $w = D_1(2R_oR_c).R_c$(say $R_c = |E_c|$) are the modules of the normalized structure factors available after the inversion of a small percentage (3.5%) of the electron density map. $\left\langle R_c^2 \right\rangle$ is calculated over the 30% of the measured reflections, those with the weakest $|F_O|$ values (they are never actively used in the phasing process);

b) for large molecules and proteins $eFOM \equiv weFOM$, as defined by Burla *et al*. (2004b):

$$weFOM = \begin{cases} \dfrac{we_1 - < we_1 >}{\sigma_1} \\ \dfrac{we_2 - < we_2 >}{\sigma_2} \\ \dfrac{we_3 - < we_3 >}{\sigma_3} \end{cases}$$

where

$$we_1 = \frac{\langle wR_c \rangle_{large}}{\langle wR_c \rangle_{weak}} \quad if \quad Rcr_{weak} < \langle Rcr_{weak} \rangle \quad and \quad Rcr_{large} > \langle Rcr_{large} \rangle$$

$$we_2 = \frac{1}{\langle wR_c \rangle_{weak}} \quad if \quad Rcr_{weak} > \langle Rcr_{weak} \rangle$$

$$we_3 = \langle wR_c \rangle_{large} \quad if \quad Rcr_{weak} < \langle Rcr_{weak} \rangle \quad and \quad Rcr_{large} < \langle Rcr_{large} \rangle$$

$\langle wR_c \rangle_{large}$ and $\langle wR_c \rangle_{weak}$ are the average values calculated for the large and the weak reflections for a given trial, $\langle Rcr_{weak} \rangle$ and $\langle Rcr_{large} \rangle$ are the average values of the crystallographic residuals obtained over all the trials of the considered block, and σ_1, σ_2, σ_3 are the corresponding standard deviations.

eFOM is expected to be a maximum for the more promising phase sets (those to be processed by DSR procedures).

The RELAX procedure

A systematic error in the ST phasing process, as in any other phasing tool, may provide well oriented but misplaced molecular fragments. *eFOM* usually ranks such trials among the most promising ones. In order to recover the information related to a translated model, it is necessary to shift the model to the correct position or, equivalently, to find the correct phase shift. The success has been obtained by the RELAX procedure (Burla *et al*., 2002). Its main steps are:

a) the so-called Cheshire cell (Hirshfeld, 1968) is automatically defined. The search of a suitable origin translation may be restricted to it.

b) The reflections are expanded in P1 (Sheldrick & Gould, 1995), in order to relax symmetry constraints. Symmetry related reflections are given values in accordance with the rule:

$$|F_{\mathbf{hR}}| = |F_{\mathbf{h}}| \qquad \varphi(\mathbf{hR}) = \varphi(\mathbf{h}) - 2\pi\mathbf{hT}\,.$$

c) Suitable figures of merit (*fom1* and *fom2*) are automatically computed by the program. The grid point X_j for which *fom1* + *fom2* is a maximum should define the correct origin translation.

d) Let X_0 be the correct origin shift: the P1 phased reflections are then modified into

$$\varphi'_{\mathbf{h}} = \varphi_{\mathbf{h}} - 2\pi\mathbf{h}\mathbf{X}_j$$

in order to turn back to the original space group in an automatic way. At this step we have re-established the original space group symmetry: to fully accomplish this task we have to select unique reflections and to assign suitable phases to them.

A default run of Sir2004 automatically applies the RELAX procedure only for few (best-ranked) trials, but the user can modify this choice.

The Patterson deconvolution procedure

Procedures alternative to the application of the tangent formula are the Patterson based methods (Bürger, 1959; Richardson & Jacobson, 1987; Sheldrick, 1992). Sir2004 uses the approach described by Burla *et al*. (2004a), which can be summarized as follows:

a) the superposition minimum function (Pavelčik, 1988; Pavelčik *et al*., 1992) is calculated by combining all the independent Harker domains of the Patterson map (whose number is denoted by $\overline{m}$) as follows:

$$SMF(\mathbf{r}) = \underset{s=1}{\overset{\overline{m}}{Min}}\left[P(\mathbf{r} - \mathbf{C}_s\mathbf{r})\right]$$

where $\mathbf{r} - \mathbf{C}_s\mathbf{r}$ is a Harker vector corresponding to the s^{th} symmetry operator and P denotes the Patterson map;

b) the minimum superposition function is obtained by:

$$S(\mathbf{r}) = Min\left[P(\mathbf{r} - \mathbf{r}_p), SMF(\mathbf{r})\right]$$

where $\mathbf{r}_p$ is the position of a pivot peak selected by the program in the SMF map;

c) filtering algorithms are applied to break, in the S($\mathbf{r}$) map, the residual Patterson symmetry;

d) the final S($\mathbf{r}$) map is inverted to provide phases and weights to start the DSR process;

The multisolution approach is obtained by using as pivots the highest SMF peaks: for each pivot a set of phases is obtained which is ranked by a specific early FOM (*pFOM*) defined as:

$$pFOM = eFOM + \frac{\int S_1(\mathbf{r})d\mathbf{r} \int S_2(\mathbf{r})d\mathbf{r}}{\left(\int P(\mathbf{r})d\mathbf{r}\right)^2}$$

where S_1 and S_2 denote the S map before and after the filtering procedure respectively, the numerator integrals are calculated for the top 2% part of the map, while the integral at denominator, representing the normalization factor, is calculated for the top 20% part of the Patterson map. It is expected that S maps having more pronounced peaks correspond to good solutions.

The Direct Space Refinement (DSR)

The DSR procedures (see Figs.2-3) are constituted by the following steps (Burla, Carrozzini, Caliandro et al., 2003):

a) s supercycles of electron density modification (EDM), each constituted by t microcycles $\rho \rightarrow \{\varphi\} \rightarrow \rho$. The default values of s and t change with the category of the structure. The modification of the electron density map includes powering (Refaat & Woolfson, 1993) and the inversion of small negative domains (Burla, Carrozzini, Caliandro *et al*., 2003). New phases and normalized structure factor modules (R_C) are obtained by inversion of a small percentage (few percents) of the electron density map: such modules are rescaled by histogram matching with respect to the distribution of the observed ones (R_O). New phases are given a Sim-like weight $w = g\left(RES/RES_{\max}\right) \times D_1\left(2R_o R_c\right)$. The function g leaves unchanged the weights for the lowest resolution reflections and smoothly increases with sinθ/λ. This feature helps the phasing process for data sets with resolutions worse than 1.2 Å.

b) When protein data are processed, the molecular envelope is calculated from the current phases. The electron density map is modified by assigning different weights to pixels falling inside or outside the envelope, so tentatively depleting the intensities of the false peaks.

c) DSR includes also cycles of HAFR (a selected number of large-intensity electron density peaks are expressed in terms of the heaviest atomic species and of suitable occupancy factors), and LSQH (the isotropic displacement parameters of the heavy atoms are refined via a least squares procedure). The reader is referred to Burla *et al*. (2001), for details. For small/medium size molecules an automatic Diagonal Least-Squares refinement (DLSQ) is applied. DLSQ is a procedure which alternates least-squares cycles and ($2|F_o|$ - $|F_c|$) map calculations, in order to complete the crystal structure, reject false peaks and refine structural parameters (Altomare *et al*., 1993). An isotropic diagonal matrix refinement is used (H atoms are not involved).

d) The whole DSR procedure could automatically be iterated for the same trial, restarting each time from the current phases (Burla, Carrozzini, Cascarano *et al*., 2003) (in the following, this technique will be called *iteration*). This iterative process, although time consuming, allows to solve also resistant molecules (i.e. protein structures diffracting at non-atomic resolution).

The molecular envelope

The molecular envelope of the protein (Wang, 1985; Leslie, 1987) is used, in Sir2004, as a mask in the density modification step, in order to improve its efficiency in solving protein structures (Burla, Carrozzini, Caliandro *et al*., 2003).

The protein volume is calculated through the Mathews (1968) formula and the envelope is calculated for each trial solution from the current phases. The electron density map is modified by assigning weights equal to 1.0 to pixels belonging to the envelope and weights equal to 0.5 to pixels out of it, so tentatively depleting the intensities of the false peaks. The map is then inverted and the resulting phases may improve their values.

The envelope information cannot be used just after the tangent formula, when a few low-resolution reflections are usually phased and when the mean phase error is normally too large. The molecular envelope is thus calculated for the first time after three macrocycles of EDM and then recursively calculated and applied in the following DSR procedures.

Identification of the correct solution: the final Fom (fFOM)

For small/medium size structures the correctness of a solution is assessed, at the end of the DSR process, by the crystallographic residual factor (*Rcr*): if the final value of *Rcr* is smaller than a given threshold (default value = 0.25) the program stops; otherwise, the program explores the next ranked phase set.

For large size molecules and proteins the least squares are very time consuming: furthermore they cannot be applied to non-atomic resolution data. To recognise the correct solution we have devised a new figure of merit (*fFOM*), to be applied at the end of the DSR process. It is defined as follows:

$$fFOM = \frac{RAT_{final}}{RAT_{initial}} \times \frac{CC(all)_{final}}{CC(all)_{initial}} \times \left(COMB_{final} - COMB_{initial}\right)$$

where CC is the correlation factor between the R_O and R_C, and

$$COMB = CC(large) + 3\ CC(weak)\ .$$

The words "*all*", "*large*" and "*weak*" indicate the complete normalized structure factors set (ranked in decreasing order), the subset of the largest $|F_o|$'s (70% of them) and the subset of the weakest ones (30% of $|F_o|$'s), respectively. The indexes "*initial*" and "*final*" indicate that the corresponding FOM values are calculated at the beginning and at the end of the DSR procedure. *fFOM* shows some quite interesting new features: a) *COMB* is constituted by two contributions: the first arising from the large and the second from the weak reflections, this last with a weight three times larger than the former. This weighting criterion is justified by the fact that the weak reflections are not

involved in the phasing process, so that *CC*(*weak*) is like a *free* index, much more reliable than the companion contribution *CC*(*large*). It is frequently negative for the wrong trials. b) *fFOM* is the product of three figures of merit: every value depends on how each FOM is modified during the DSR process, not on the value it assumes at the end of the process.

The correct solution should be identified by large values of *fFOM*.

Sir2004 strategy for ab initio phasing of crystal structures

The Sir2004 flow diagram, shown in Fig.1, is a useful guide for understanding the program strategy. We note :

a) The tangent formula is particularly efficient for small structures: here the role of the DSR is quite marginal. For medium size structures the application of the DSR is more important: it is often able to drive phases with large errors to their correct values. For proteins the tangent procedures are rather inefficient: a large set of trials have to be explored before finding the useful one. The Patterson techniques recently developed by Burla *et al.* (2004a) are frequently able to find the correct solution in few trials, provided some heavy atoms are present in the structure. In our experience, if the solution is not found in the first Patterson-derived trials (obtained by using the largest SMF peaks as pivots), it is unlikely to find the correct solution in the following ones. In accordance with such conclusions SIR2004 uses the following strategy :

i) by default, for small and medium size structures, only the ST algorithm is used;

ii) for large structures the phasing process starts with the Patterson procedure, except for the following cases: absence of heavy atoms (Z_H<11), very large structures (more than 1000 atoms in the asymmetric unit) characterized both by data resolution worse than 1.2Å and by intermediate heavy atoms (up to Ca); in these cases the ST is directly used. Only 30 Patterson phase sets are generated and explored by the DSR module.

Three blocks of ST trials are subsequently generated. The trials in each block (let n_i be their number for the i^{th} block) are sorted by *eFOM* and the DSR module is applied only to the subset of best ranked trials (let m_i be their number). The values of n_i and m_i are automatically settled by the program, as a function of the complexity of the structure, in accordance with the defaults schematised in Table 1. Default values of *n* and *m* for proteins (*n* and *m* do not change with the block number) are shown in Table 2: the data resolution is also taken into account (proteins data collected at non-atomic resolution are more difficult to phase). By means of directives the user can modify the default choices (Tables 1 and 2 are useful to guide the user towards sensitive non-default values; *Nasym* is the number of non-hydrogen atoms in the asymmetric unit, Z_H is the atomic number of the heaviest species).

b) The DSR module (see Figs.2-3) is constituted by cycles of EDM (electron density modification), HAFR (a selected number of large-intensity electron density peaks are expressed in terms of the heaviest atomic species and of suitable occupancy factors), and LSQH (the isotropic displacement parameters of the heavy atoms are refined *via* a least squares procedure). The reader is referred to Burla *et al*. (2002), for details. According to the structural complexity, the strategy and the algorithms improved in the DSR section are different (see Table 3).

The number of total DSR *iterations* (default mode) is automatically defined by the program and it is shown in Table 4. It is a function of *Nasym*, RES and Z_H. The user can change the number of the iterations either for saving computing time or for increasing the chances to solve resistant structures.

The RELAX procedure is performed only on the first n_R ST trials (the best *eFOM*-ranked), where n_R is set to 3, 5 or 20 for small, medium and large size molecules, respectively. This choice is due to two reasons: i) the use of RELAX is time consuming; ii) our tests suggest that trials corresponding to well oriented but misplaced molecules are usually characterized by large values of *eFOM*. RELAX is not applied to Patterson phase sets.

For small and medium size structures the program stops if the final value of *Rcr* is smaller than a given threshold (default value = 0.25), while, for macromolecules, the PHASE module runs until all the trials fixed by default or by the user have been processed. A histogram is continuously updated showing the *fFOM* distribution. If, for a given trial, *fFOM* is outstanding, the user can stop the program and submit the selected trial to further analysis (i.e., examination of the electron density map by means of contouring facilities, automatic model building, ….) using the Final Stage Procedures (accessible through the graphical interface). If the user is unsatisfied, he can launch the PHASE module again, starting from the examined trial. The above strategy allows to save computing time if the user looks at the *fFOM* histogram.

Sir2004: Completion and refinement of the structure

The least squares refinement

As mentioned before, for small/medium size molecules, a preliminary diagonal least squares refinement is automatically performed at the end of the DSR process to recognize the correct solution. The least squares module is also suitable for the complete crystal structure refinement. Its specific features are:

1) the full matrix may be used or any kind of blocks.
2) 18 weighting schemes are available. If the weighting scheme contains adjustable parameters, the program refines the values to obtain a good distribution of $<w\Delta^2>$ against $|F_o|$ and resolution, and the value of the goodness of fit close to one (Spagna & Camalli, 1999).
3) The program generates constraints for the parameters of atoms on special positions in all space groups.
4) Automatic or through wizard generation of hydrogen atoms. Their contributions are included in the refinement, by allowing the positional parameters to ride on the corresponding parent atom.
5) Floating origin is restrained automatically by setting the restrain on the sum of the appropriate coordinates.
6) Refinement of the Flack parameter to evaluate the absolute configuration.
7) The possibility to impose conditions (constraints) or additional information (restraints).

The constrained atoms are regularized to an ideal model structure of known geometry (i.e. benzene ring) and this rigid body is refined as compact unit assuming three translational parameters and three angles which define its orientation. The method used to compute the coordinates of the model follows the approach described by Arnott & Wonacott (1966). In order to build the internal Cartesian coordinates, the program uses the ASCII file Sir2004.gru which contains models described by the Z-matrix formalism.

The following restraints are available: bond distances, bond angles, planarity.

Fourier, least squares, hydrogens and restraints tools are accessible through the Graphical User Interface (GUI).

When default Sir2004 fails: some advices

Sir2004 has developed an automatic strategy to find the correct solution among the various trials. In addition the user can adopt several options to choose his own phasing pathway. We quote:

a) the value of NREF (number of reflections actively used in the phasing process ≡ N_{large}) is fixed by the program. For some special structures the ratio "number of active triplets/NREF" is too small (e.g., less than 20). Larger values of NREF may improve the phasing procedure.

b) Because of a faulty data collection strategy, often weak reflections may not be included in diffraction data. This lack of information influences both the normalization process (scale and overall thermal factors are affected by systematic errors; the experimental E-distribution is often non-centric even when the crystal structure is centrosymmetric) and the estimation of invariants: in particular, a reduced number of negative triplets (*via* P_{10} formula) and of negative quartets is calculated. Success in the structure solution may be obtained if weak reflections are also used.

c) High (or low) resolution reflections may occasionally play a too important role in the first steps of the phasing process. Fixing a thermal factor lower (or larger) than that provided by the normalization routine may successfullly change the phase extension and refinement procedures.

d) Use of the RELAX procedure.

e) An alternative space group should be carefully considered.

Fig. 1 – Flow diagram of the SIR2004 program.

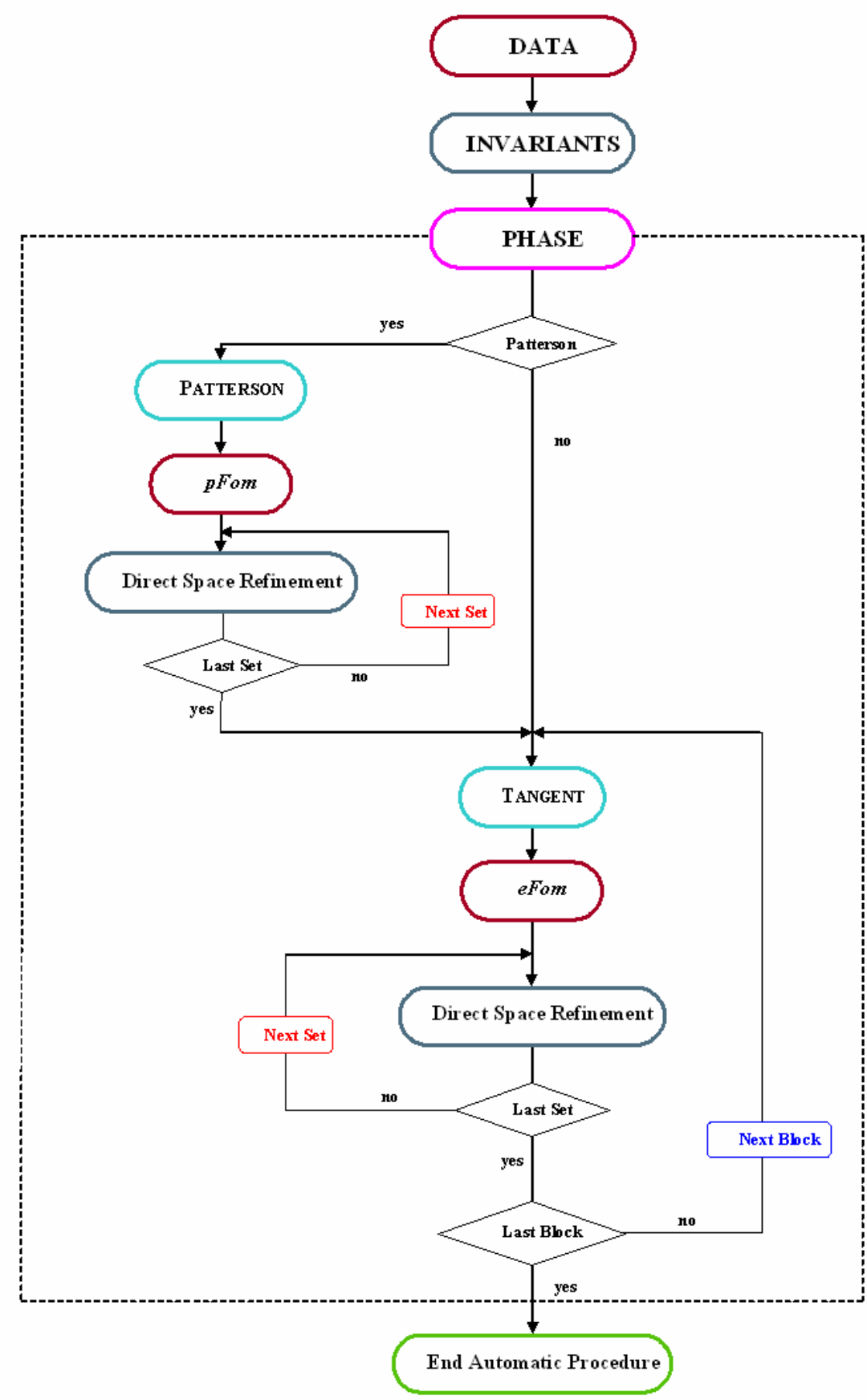

Fig. 2 – Flow diagram of the DSR procedure for small and medium molecules. *Rcr* is the crystallographic residual factor and *Rth* is its related threshold (in default *Rth* = 25%).

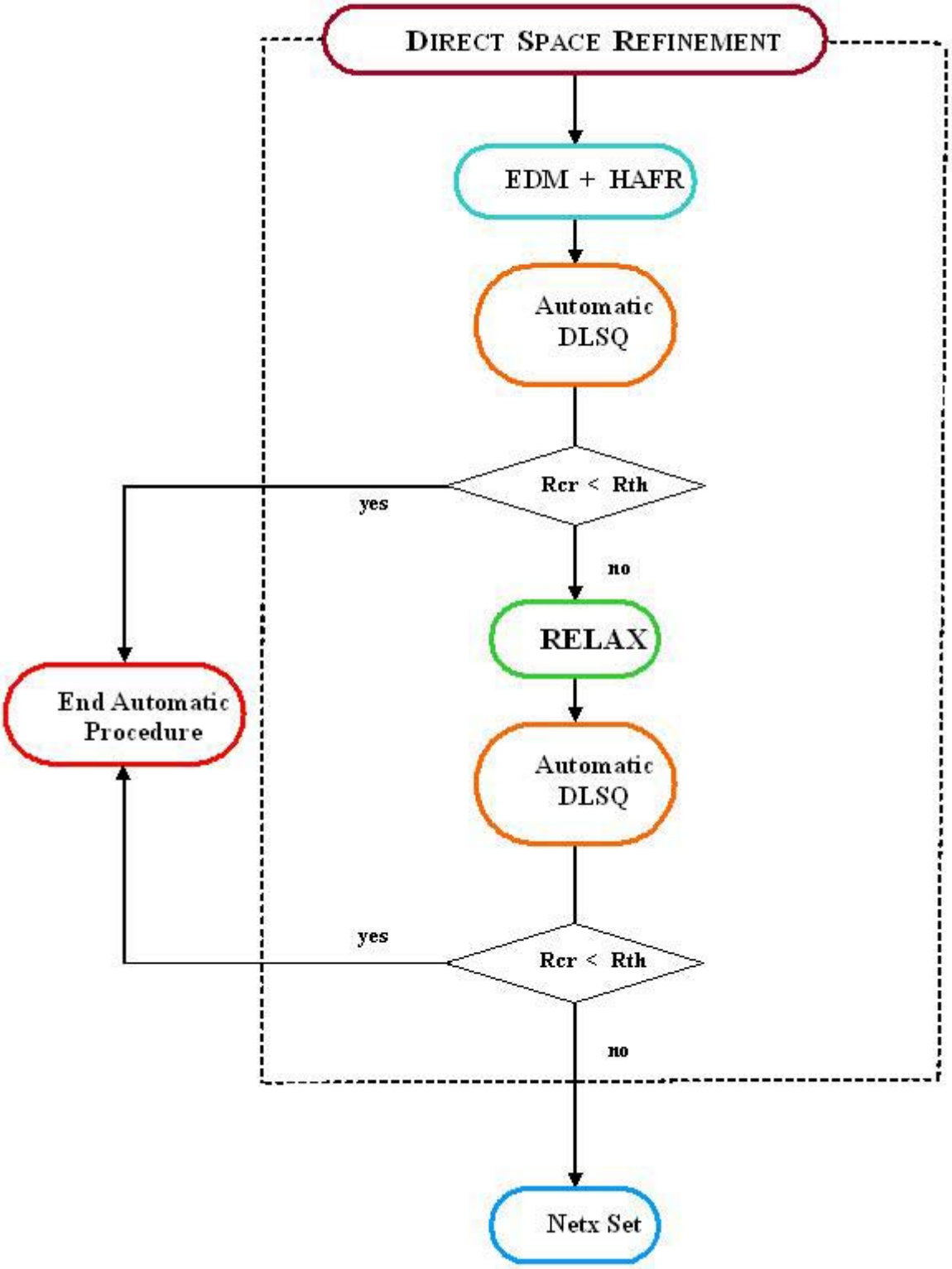

Fig. 3 - Flow diagram of the DSR procedure for macromolecules.

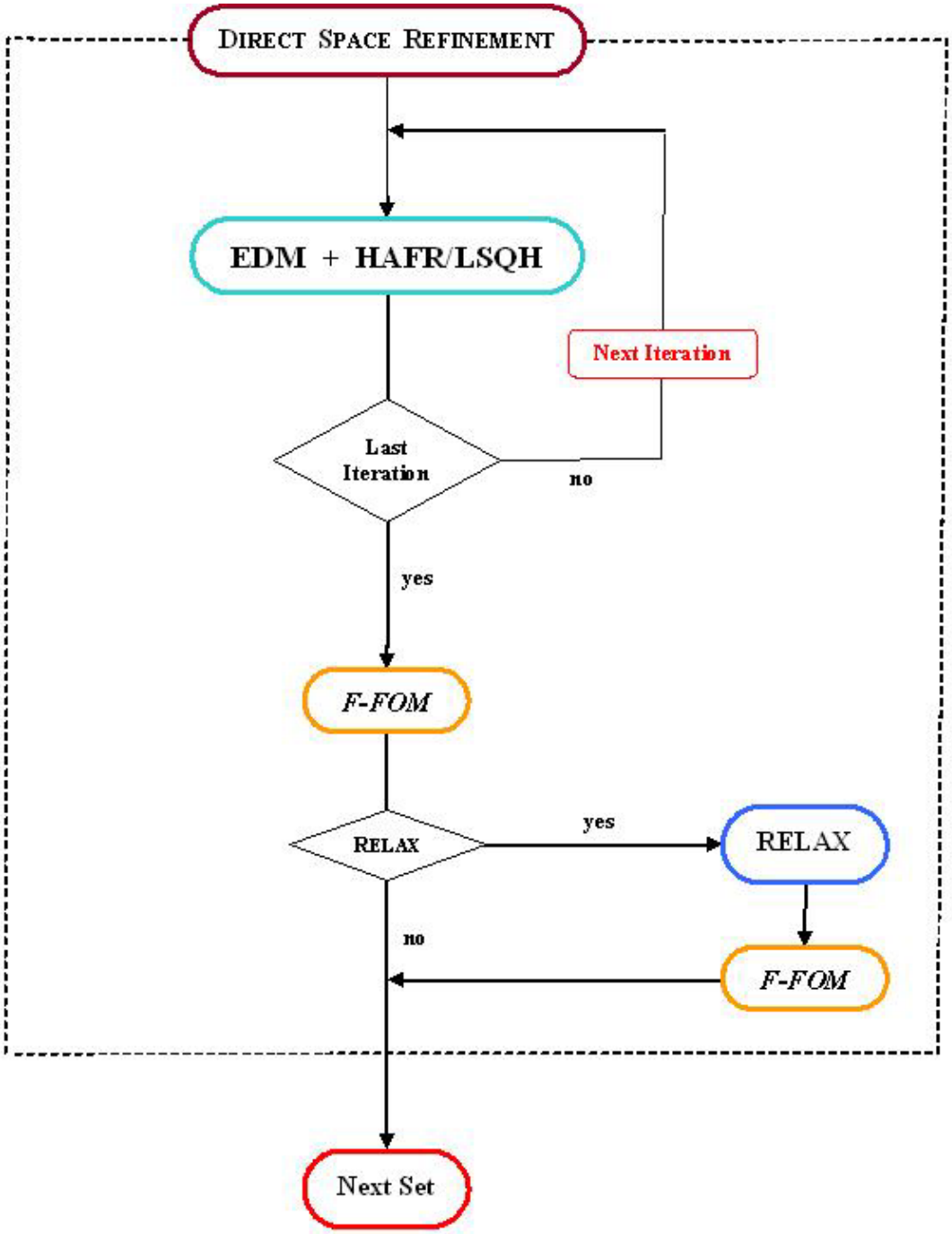

Table 1. Small and medium size molecules: default values of (n_i, m_i) for each block of trial solutions.

	1st Block	**2nd Block**	**3rd Block**
Nasym < 81	(100, 10)	(100, 20)	(100, 30)
80 < *Nasym* < 201	(300, 50)	(300, 100)	(300, 200)

Table 2. Large structures **:** default values of (*n, m*) versus structure complexity and data resolution.

	ATOMIC RES $Z_H > 20$	**ATOMIC RES $12 < Z_H < 21$**	**NON-ATOMIC RES $Z_H > 20$**	**NON-ATOMIC RES $12 < Z_H < 21$**	**ANY RES $Z_H < 12$**
200 < *Nasym* < 601	(200, 100)	(400, 200)	(200, 100)	(600, 300)	(2500, 300)
600 < *Nasym* < 1001	(200, 100)	(1200, 300)	(1500, 300)	(2100, 300)	(2500, 300)
Nasym > 1000	(2500, 300)	(2500, 300)	(2500, 300)	(2500, 300)	(2500, 300)

Table 3. Strategy of DSR procedures as function of structural complexity.

Size	Atoms in a.u.	EDM	HAFR	LSQH
XS	up to 5	S = 6 ÷ M = 5	N = 7	NO
S	6 - 80	S = 6 ÷ M = 5	N = 7	NO
M	81 - 200	S = 9 ÷ M = 5	N = 7	NO
L	201 - 600	S = 15 ÷ M = 7	N = 6	Heavy atoms
XL	601 - 1000	S = 15 ÷ M = 7	N = 6	Heavy atoms
XXL	more the 1000	S = 15 ÷ M = 7	N = 6	Heavy atoms

Table 4. Default number of DSR iterations *versus Nasym*, *RES* and Z_H .

	ATOMIC RES $Z_H > 20$	**ATOMIC RES** $Z_H < 21$	**NON-ATOMIC RES** $Z_H > 20$	**NON-ATOMIC RES** $Z_H < 21$
Nasym < 81	---	---	---	---
80 < *Nasym* < 201	---	---	---	---
200 < *Nasym* < 601	---	---	---	5
600 < *Nasym* < 1001	1	4	7	7
Nasym > 1000	1	4	9	9

Commands and their use

The input consists of a sequence of comments, commands and directives. The commands are headed by '%' character and directives must follow the related command.

Sir2004 recognizes the following commands:

%INITIALIZE Initialize the direct access file (to override previous results and data)

%DATA Data input routine

%INVARIANTS Invariants routine

%PHASE Phasing routine + Direct Space Refinement routine

%END End of the input file

%JOB A caption is printed in the output

%CONTINUE The program runs in default conditions from the last given command up to the end

%STRUCTURE *string* This command is used to specify the name of the structure to investigate.

The program creates the name of some needed files, adding the appropriate extension to the structure name. The file names are:

string.bin -> direct access file

string.res -> file in SHELX frmat

string.plt -> file for graphics

If STRUCTURE command is not used the default string "STRUCT" (instead of the name of the structure) is used to create file names.

%WINDOW Graphic window is required.

%NOWINDOW Graphic window is suppressed

Directives are described below, in the sections dedicated to the various routines.

All commands and directives are in free format (between columns 1-80) and are case independent. Only the first four characters are significant. The keywords can start in any position. If the first non-blank character is ">" or "!" then the record is interpreted as a comment; characters following ">" or "!" will be ignored.

Sir2004 preserves intermediate results. For example, if invariant estimates have been already obtained during a previous run of Sir2004, in a new run the commands %INVARIANTS can be omitted.

Commands can be given in any order, under the following conditions:

- first routine used must be DATA, if it has not been used in a previous run;
- INVARIANTS routine has no meaning if observed structure factors are not normalized;
- PHASE routine has no meaning if no triplets have been calculated.

The minimal information needed by Sir2004 is constituted by:

- cell parameters
- cell content
- space group symbol
- reflections

Directives and their use

Data Routine

(Only directives marked in **red** are mandatory.

CELL a b c α β γ

Cell dimensions a, b, and c are in Ångstrom, α, β and γ in degrees.

ERRORS esd(a) esd(b) esd(c) esd(α) esd(β) esd(γ)

Estimated standard deviations for the unit cell dimensions.

SPACEGROUP *string*

String is the symbol or the number of the space group, according to International Tables (1974). Blanks are necessary among the terms constituting the space group symbol (see examples at the end of this manual).

CONTENT El_1 n_1 El_2 n_2 El_3 n_3 .....

Unit cell content. El_i is the chemical symbol of atomic type *i*, n_i is the corresponding number of atoms in the unit cell (up to a maximum of 8 atomic types). For each chemical element up to Californium (Cf, Z=98) X-ray and electrons scattering factor constants are stored, together with information on the atomic number and weight, covalent and Van-der-Waals radii, etc. in a file (see notes on implementation).

SFACTORS El a_1 b_1 a_2 b_2 a_3 b_3 a_4 b_4 c

Scattering factors for species El. If more lines are necessary, use character = at the end of the line.

ANOMALOUS El Δf' Δf"

Values of Δf' and Δf" for species El.

RHOMAX x

Maximum value of $\sin(\Theta/\lambda)^2$ accepted for reflections to be used. In default all the data are accepted.

RESMAX x

Maximum value of resolution (in Ångstrom) accepted for reflections to be used. In default all the data are accepted.

FORMAT *string*

String is the run time format to read reflections. Default value for *string* is (3I4,2F8.2).

RECORD n

It specificies the number of reflections per record, when n>1.

REFLECTIONS *string* [$R_{1,1}$ $R_{1,2}$ $R_{1,3}$ $R_{2,1}$ $R_{2,2}$ $R_{2,3}$ $R_{3,1}$ $R_{3,2}$ $R_{3,3}$]

String is the name of the reflections file. Records have n reflections, each with h,k,l, |F|, σ(F) where h,k,l are integer (up to 512). If the orientation matrix $R_{i,j}$ is supplied, it is immediately applied to reflections and all calculations will be performed using the final orientation. The end of reflections is detected using one of the following:

- blank record.
- end of file.

Negative values of |F| are allowed; negative values of σ(F) are forbidden.

NOSIGMA

To be used when σ(F)values are meaningless or not available.

FOBS

Program assumes h,k,l, F, σ(F). F^2 and σ(F^2) are expected as default.

FOSQUARED

Program assumes h,k,l, F^2 and σ(F^2) (default choice).

WAVE *string*

Nevertheless the wavelength is not used during the structure solution stage, it is necessary for the LSQ-refinement; its value is also written in CIF file produced by the program. Possible values for *string* are: Cu, Mo or a numeric value. Default wavelength is Mo.

NREFLECTION n

Number of active reflections with largest |E| values (up to 4000) subject to a minimum value of |E| = 1.2. Default number is computed by the program.

BFACTOR x

Isotropic thermal factor, if the user wants to supply it. (The scale factor is assumed equal to 1).

ELECTRONS

This directive specifies that electron diffraction data will be used.

Invariants Routine

GMIN x

Triplets with G < x are not actively used. Default value x = 0.3 (in any case x > 0.1).

COCHRAN

To use the Cochran distribution. (P_{10} formula is used by default).

NQUARTETS

To not calculate negative quartets (Default uses them).

Phase Routine

SIZE xs, s, m, l, xl, xxl

To set a suitable solution strategy. Default procedure is automatically chosen by the program, on the basis of the structural complexity.

NNEG

The program does not actively use negative triplet relationships.

NNQG

The program does not actively use negative quartet relationships.

BLOCK m n_{11} n_{12} ….. n_{i1} n_{i2}

The program explores m blocks (up to 10) of n_{i1} trials and then selects the most promising n_{i2} (up to 300) phase sets (those with the higher *eFOM* score) to perform DSR procedures. Default values are automatically chosen by the program, on the basis of the structural complexity.

ITERATION n

The program performs n (up to 25) Direct Space Refinement iteration(s).

TRIAL n

The program explores only the tangent trial associated with the progressive number n.

STRIAL n

The program starts from the n^{th} tangent trial.

RELAX

The programs applies the RELAX procedure, exploring all the phase sets selected for any block of trials.

UNRELAX

The programs does not apply the RELAX procedure.

PATTERSON

The program applies the Patterson procedure.

NOPATTERSON

The program does not apply the Patterson procedure.

PEAKS n

The program applies the Patterson procedure to the n (up to 50) highest peaks in the SMF map.

PTRIAL n

The program explores only the Patterson trial associated with the peak number n.

NOLSQ

The program does not perform the automatic Diagonal Least-Squares calculations.

CYCLE n

The program stops at cycle n of the automatic Diagonal Least-Squares calculations.

RESIDUAL x

The program stops if the final crystallographic residual factor (Rcr %) is less than the specified value x. The default value is 25%. This directive is meaningless for macromolecules.

FRAGMENT *string*

Used to supply a known fragment. *String* is the name of the file in which, for each atom, are stored the following data: Element X Y Z B(iso).

RECYCLE

Used to complete a known fragment supplied to Sir2004.

CRYSTALS

The user wants to produce the output file in CRYSTALS format (SHELX format is used by default).

Examples of input for Sir2004

Example 1

The following example shows the maximum default use of Sir2004. Most of the structures can be solved in this way. Diffraction data are in the file crambin.hkl, in format (3I4,2F8.2), one reflection per record.

```
%Data
    Cell   40.763 18.492 22.333 90.00 90.61 90.00
    SpaceGroup  P 21
    Content    C 406 H 776 N 110 O 131 S 12
    Reflections  crambin.hkl
%Continue
```

Example 2

In the following example, experimental data are stored as |F| (not $|F|^2$), using the format (3(3i4,f10.3,8x,f8.2)), 3 reflections per record.

```
%Window
%Structure    iled
%Job          Isoleucinomycin
%Initialize
%Data
    cell             11.516  15.705  39.310  90.00  90.00  90.00
    spacegroup       P 21 21 21
    content          C 240  H 408  N 24  O 72
    reflections      iled.hkl
    record           3
    format           (3 (3i4,f10.3,8x,f8.2) )
    fobs
%Continue
```

Example 3

The user wants to supply the value for the isotropic thermal factor and to set the number of strong (|E| value) reflections .

```
%Window
%Structure ferre
%Job    ferredoxin (pdb code: 2fdn)
%Initialize
%Data
      cell            33.95  33.95  74.82  90.00  90.00  90.00
      spacegroup      P 43 21 2
      content         C 1824  H 2744  N 488  O 478  S 128  Fe 64
      reflections      ferre.hkl
      bfac            3.5
      nref            2000
%Continue
```

Example 4

In the following example, the Cochran formula is applied and all triplets with a concentration parameter greater than 0.2 are actively used in the phasing process, as requested by the user. The binary file "ferre.bin" must exist. Commands or directives following ">" or "!" are interpreted as a comment and will be ignored.

```
%WINDOW
%STRUCTURE ferre
>%INITIALIZE
>%DATA
      > CELL 33.95 33.95 74.82  90.00 90.00 90.00
      ! SPACEGROUP P 43 21 2
      ! CONTENT C 1824  H 2744  N 488  O 478  S 128  Fe 64
      > REFLECTIONS ferre.hkl
%INVARIANTS
      GMIN 0.2
      COCHRAN
%PHASE
%END
```

Example 5

The user wants to explore new trials, starting from the trial number 132, from the phasing process up to the Fourier-Least Squares refinement. The graphical interface is not used. The binary file "iled.bin" must exist.

```
%Nowindow
%Structure iled
%Phase
   strial  132
%End
```

Example 6

The user wants to explore only trial number 154. The program stops before DLSQ calculations.

```
%Nowindow
 %Structure loganin
%Phase
      trial 154
      Nolsq
%End
```

Example 7

The user wants to explore 2 blocks of 300 trials, storing 50 sets for the first and 100 sets for the second; the program does not apply the RELAX procedure.

```
%Window
 %Structure crambin
%Phase
      Blocks 2  300 50  300 100
      Unrelax
%End
```

Example 8

The program explores only the Patterson trial associated with the peak number 5; it applies the Direct Space Refinement strategy for medium size structures and stops if the final residual value (*Rcr* %) is less than the threshold indicated by the user.

```
%Nowindow
 %Structure conotoxin
%Phase
      Ptrial 5
      Residual 10.0
      Size m
%End
```

Example 9

The user wants to apply the Patterson procedure to the 25 highest peaks in the SMF map.

```
%window
 %Structure conotoxin
%Phase
      Peaks 25
%End
```

Example 10

The user wants to explore only the tangent trial associated with the progressive number 26 and stop the program at cycle 8 of the DLSQ procedure.

```
%window
 %Structure loganin
%Phase
      trial 26
      cycle  8
%End
```

Example 11

In the following example the user knows a fragment and wants to complete it using the Fourier-Least Squares procedure. The ascii file "azet.fra" must exist.

```
%WINDOW
%STRUCUTRE  azet
%Phase
  Fragment   azet.fra
  Recycle
%CONTINUE
```

Coordinates are in the file "azet.fra" which contains

```
Cl    .02944    .72012    .08865
Cl    .23727    .78692    .30869
```

References

Altomare, A., Cascarano, G., Giacovazzo, C. & Guagliardi, A. (1993), *J. Appl. Cryst.*, **26**, 343-350.

Arnott, S.& Wonacott, A.J. (1966). *Polymer*, **7**, 157-166.

Baggio, R., Woolfson, M.M., Declerq, J.P. & Germain, G.(1978), *Acta Cryst.*, A**34**, 883-892.

Burla, M.C., Caliandro, R., Camalli, M., Carrozzini, B., Cascarano, G., De Caro L., Giacovazzo, C. Polidori, G. & Spagna, R.(2004), *submitted*.

Burla, M.C., Caliandro, R., Carrozzini, B., Cascarano, G., De Caro L., Giacovazzo, C. & Polidori, G. (2004a), *J. Appl. Cryst.*, **37**, 258-264.

Burla, M.C., Caliandro, R., Carrozzini, B., Cascarano, G., De Caro L., Giacovazzo, C. & Polidori, G. (2004b), *J. Appl. Cryst.*, **37**, 791-801.

Burla, M.C., Camalli, M., Carrozzini, B., Cascarano, G., Giacovazzo, C., Polidori, G. & Spagna, R. (2000), *Acta Cryst.*, A**56**, 451-457.

Burla, M.C., Camalli, M., Carrozzini, B., Cascarano, G., Giacovazzo, C., Polidori, G. & Spagna, R. (2001), *J. Appl. Cryst.*, **34**, 523-526.

Burla, M.C., Camalli, M., Carrozzini, B., Cascarano, G., Giacovazzo, C., Polidori, G. & Spagna, R. (2003), *J. Appl. Cryst.*, **36**, 1103.

Burla, M.C., Carrozzini, B., Caliandro, R., Cascarano, G., De Caro L., Giacovazzo, C. & Polidori, G. (2003), *Acta Cryst.*, A**59**, 560-568.

Burla, M.C., Carrozzini, B., Cascarano, G., De Caro, L., Giacovazzo, C., & Polidori, G. (2003), *Acta Cryst.*, A**59**, 245-249.

Burla, M.C., Carrozzini, B., Cascarano, G., Giacovazzo, C. & Polidori, G. (2002), *Z. Kristallogr.*, **217**, 629-635.

Burla, M.C., Cascarano, G. & Giacovazzo, C. (1992), *Acta Cryst.*, A**48**, 906-912.

Bürger, M. J. (1959). *Vector Space*, chapter 11. Wiley, New York.

Burzlaff, H. & Hountas, A. (1982), *J. Appl. Cryst.*, **15**, 464-467.

Cascarano,G., Giacovazzo, C. & Guagliardi, A. (1991), *Acta Cryst.*, A**47**, 698-702.

Cascarano, G., Giacovazzo, C. & Luić, M. (1988 a), *Acta Cryst.*, A**44**, 176-183.

Cascarano, G., Giacovazzo, C. & Luić, M. (1988 b), *Acta Cryst.*, A**44**, 183-188.

Cascarano, G., Giacovazzo, C., Burla, M.C., Nunzi, A. & Polidori,G. (1984), *Acta Cryst.*, A**40**, 389-394.

Cochran, W.(1955), *Acta Cryst.*, **8**, 473-478.

Fan, Hai-Fu, Yao, Jia-Xing & Qian, Jin-Zi (1988), *Acta Cryst.*, A**44**, 688-691.

Giacovazzo, C. (1976), *Acta Cryst.*, A**32**, 958-966.

Giacovazzo, C. (1977), *Acta Cryst*. A**33**, 933-944.

Giacovazzo, C. (1980), *Acta Cryst*. A**36**, 362-372.

Giacovazzo, C., Burla, M.C. & Cascarano, G. (1992), *Acta Cryst*., A**48**, 901-906.

Hirshfeld, F. L. (1968), *Acta Cryst*., A**24**, 301-311.

Leslie, A. G. W. (1987), *Acta Cryst*., A**43**, 134-136.

Matheus, B.W. (1968), *J. Mol. Biol*., **33**, 491-497.

Pavelčik, F. (1988), *Acta Cryst*. A**44**, 724-729.

Pavelčik, F., Kuchta, L. & Sivy, J. (1992), *Acta Cryst*. A**48**, 791-796.

Refaat, L. S. & Woolfson, M. M. (1993), *Acta Cryst*. D**49**, 367-371.

Richardson, J.W. & Jacobson, R. A. (1987). In: *Patterson and Pattersons*. Ed. by Glusker, J. P., Patterson, B. K.&. Rossi, M, pp. 310-317. Oxford University Press.

Sheldrick, G. M. (1992). In: *Crystallographic Computing 5*. Ed. by Moras, D., Podjarny, A. D. & Thierry, J. C., pp. 145-157. Oxford University Press.

Sheldrick, G. M. & Gould, R. O (1995), *Acta Cryst*., B**51**, 423-431.

Spagna, R. & Camalli, M. (1999). *J. Appl. Cryst*., **32**, 934-942.

Tronrud, D.E. (1997), *Methods Enzymol*., **277**B, 306-319.

Wang, B. C. (1985), *Methods Enzymol*., **115**, 90-112.